

# Inexact Regularized Proximal Newton Method: Provable Convergence Guarantees for Non-Smooth Convex Minimization without Strong Convexity

Man-Chung Yue\*      Zirui Zhou†      Anthony Man-Cho So‡

May 25, 2016

## Abstract

We propose a second-order method, the *inexact regularized proximal Newton* (IRPN) method, to minimize a sum of smooth and non-smooth convex functions. We prove that the IRPN method converges globally to the set of optimal solutions and the asymptotic rate of convergence is superlinear, even when the objective function is not strongly convex. Key to our analysis is a novel usage of a certain error bound condition. We compare two empirically efficient algorithms—the *newGLMNET* [28] and adaptively restarted *FISTA* [16]—with our proposed IRPN by applying them to the  $\ell_1$ -regularized logistic regression problem. Experiment results show the superiority of our proposed algorithm.

## 1 Introduction

A wide range of tasks in machine learning and statistics can be formulated as a non-smooth convex minimization of the following form:

$$\min_{x \in \mathbb{R}^n} F(x) := f(x) + g(x), \quad (1)$$

where  $f : \mathbb{R}^n \rightarrow (-\infty, \infty]$  is convex and twice continuously differentiable on an open subset of  $\mathbb{R}^n$  containing  $\text{dom}(g)$  and  $g : \mathbb{R}^n \rightarrow (-\infty, \infty]$  is a proper, convex and closed function, but possibly non-smooth. First-order methods, such as proximal gradient algorithm and its accelerated versions, are among the most popular choices for solving problem (1) and have been the subject of intense research over the last decade. However, they suffer from two major drawbacks. First, they are particularly slow for reaching solutions with high accuracy, especially for ill-conditioned problems. Second, for problems where the function value or the gradient is expensive to compute, such as the logistic loss function and likelihood functions arising in conditional

---

\*Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E-mail: [mc Yue@se.cuhk.edu.hk](mailto:mc Yue@se.cuhk.edu.hk)

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E-mail: [zrzhou@se.cuhk.edu.hk](mailto:zrzhou@se.cuhk.edu.hk)

‡Department of Systems Engineering and Engineering Management, and, by courtesy, CUHK-BGI Innovation Institute of Trans-omics, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E-mail: [mancho so@se.cuhk.edu.hk](mailto:mancho so@se.cuhk.edu.hk)

random fields, employing second-order information can improve the overall performance of the algorithms [22, 28, 29].

In this paper, we explore a Newton-type method for solving (1): the successive quadratic approximation (SQA) method. At the iteration  $k$  of a generic SQA, one computes the minimizer  $\hat{x}$  (or an approximation of it) of a quadratic model  $q_k(x)$  of the objective function  $F(x)$  at  $x_k$ :

$$q_k(x) := f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2} (x - x_k)^T H_k (x - x_k) + g(x), \quad (2)$$

where  $H_k$  is a positive definite matrix approximating the Hessian matrix  $\nabla^2 f(x_k)$  of  $f$  at  $x_k$ . A backtracking line search with step size  $\alpha_k$  along the direction  $d_k := \hat{x} - x_k$  is then performed and  $x_{k+1} := x_k + \alpha_k d_k$  is returned as the next iterate. Since in most cases, the minimizer  $\hat{x}$  (no matter exact or approximate) of (2) does not admit a closed-form expression, an iterative algorithm, which we shall refer to as the *inner solver*, is invoked.

There are three important ingredients that, more or less, determine an SQA method: the approximate Hessian  $H_k$ , the inner solver for minimizing  $q_k$ , and the stopping criterion of the inner solver for controlling the inexactness of  $\hat{x}$  to the minimizer of  $q_k(x)$ . Indeed, many existing SQA methods and their variants that are tailored for special instances of (1) can be obtained by specifying the aforementioned ingredients. Friedman et al. [8] developed the GLMNET algorithm for solving the  $\ell_1$ -regularized logistic regression, where  $H_k$  is set to be the exact Hessian  $\nabla^2 f(x_k)$  and coordinate minimization is used as the inner solver. Yuan et al. [28] improved GLMNET by replacing  $H_k$  with  $\nabla^2 f(x_k) + \nu I$  for some constant  $\nu > 0$  and adding a heuristic adaptive stopping strategy for inner minimization. This algorithm, called newGLMNET, is now the workhorse of the well-known LIBLINEAR package [7] for large scale linear classification. Hsieh et al. [9] proposed the QUIC algorithm for solving sparse inverse covariance matrix estimation, which makes use of a quasi-Newton model for forming the quadratic approximation (2). Similar strategies have also been employed in Olsen et al. [17], where the inner problems are solved by the fast iterative soft-shrinkage algorithm (FISTA) [1]. Other SQA variants can be found, *e.g.*, in [22, 2, 29].

Although the numerical advantage of the above algorithms have been well documented, their theories on global convergence and convergence rate require that the inner problems can be solved exactly. This is rarely satisfied in practice. To address the inexactness of the inner problem, Lee et al. [10] and Byrd et al. [3] recently proposed several families of SQA methods along with stopping criteria for inner problems that yield global convergence and asymptotic superlinear rate. One crucial assumption, without which the analyses break down or the algorithms are not well-defined<sup>1</sup>, that underpins their theoretical results is the strong convexity of  $f$ . Unfortunately, it is by now well known that such an assumption does not hold in many applications of interest. For example, the objective function in the  $\ell_1$ -regularized logistic regression is

$$F(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-b_i \cdot a_i^T x)) + \mu \|x\|_1,$$

where the smooth part  $f(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-b_i \cdot a_i^T x))$  is strongly convex if and only if the data matrix  $A := [a_1, \dots, a_m] \in \mathbb{R}^{n \times m}$  is of full row rank, which is not a guaranteed property

---

<sup>1</sup>For [10],  $H_k = \nabla^2 f(x_k)$ . If strong convexity is missing, the quadratic model (2) is not strongly convex and can have multiple minimizers. Hence the next iterate  $x_{k+1}$  is not well-defined.

in most applications and is even impossible if the data collection is under the high-dimensional setting; *i.e.*,  $n \gg m$ . Therefore, the theories of existing SQA methods are rather incomplete.

Recognizing these deficiencies, we develop the *inexact regularized proximal Newton* (IRPN) method, which is an SQA method for solving (1) and has provable convergence guarantees even without exact inner solutions or the strong convexity assumption. At iteration  $k$ , IRPN takes  $H_k$  to be the regularized Hessian  $H_k = \nabla^2 f(x_k) + \mu_k I$ , where  $\mu_k = c \cdot r(x_k)^\rho$  for some constants  $c > 0$ ,  $\rho \in [0, 1]$ , and  $r(\cdot)$  is a residual function measuring how far the current iterate  $x_k$  is from the optimal set (see Section 2.2 for details). In addition, the inner minimization at iteration  $k$  is solved up to an adaptive inexactness condition that also depends on the residual  $r(x_k)$  and the constant  $\rho$ . It is worth noting that the idea of regularization is not new for second-order methods. The well-known Levenberg-Marquardt (LM) method for solving nonlinear equations is essentially a regularized Gauss-Newton method [15, 26]. Regularized Newton methods have also been proposed for solving smooth convex minimization with singular solutions [11]. The algorithm that is closest in spirit to ours is the above-mentioned **newGLMNET**. However, the regularization parameter is chosen empirically and remains constant ( $\mu_k = \nu$ ) throughout the entire execution, which limits the convergence rate of **newGLMNET** to be at most linear. Furthermore, a heuristic stopping rule is adopted for the inner minimization.

In the sequel, we analyze the global convergence as well as the local convergence rate of IRPN without assuming strong convexity. We prove that the sequence of iterates generated by IRPN converges globally, in the sense that each accumulation point of the sequence is an optimal solution to (1). Such result is true even when the sequence of regularization parameters  $\{\mu_k\}_{k \geq 0}$  diminishes to zero. By contrast, all existing results on global convergence of inexact SQA methods require the sequence of matrices  $\{H_k\}_{k \geq 0}$  to be uniformly positive definite [10, 3]; *i.e.*, there exists a constant  $\lambda > 0$  such that  $H_k \succeq \lambda I$  for all  $k$ . As for the rate of convergence, we prove that the distance of the iterates of IRPN to the set of optimal solutions converges to 0 *linearly* if  $\rho = 0$ , *superlinearly* if  $\rho \in (0, 1)$  and *quadratically* if  $\rho = 1$ , provided that certain error bound (EB) condition holds (see Section 2.2). This specific EB condition originates from Luo and Tseng [14] and has been proved to hold for a wide range of instances of (1) (see Theorem 1). Moreover, such EB condition played a fundamental role in establishing linear convergence of various first-order methods [14, 25, 24, 31]. To the best of our knowledge, it is the first time that such an EB condition is utilized for proving superlinear or quadratic convergence rate of SQA methods for problem (1).

One immediate implication of our theoretical developments is that when applied to solve  $\ell_1$ -regularized logistic regression, the distance of the iterates of IRPN to the set of optimal solutions converges to zero globally with asymptotic linear, superlinear or even quadratic convergence rate (depending on the choice of  $\rho$ ), regardless of whether the logistic loss function is strongly convex or not. We compare two widely used, efficient algorithms—the **newGLMNET** algorithm [28] and the adaptively restarted FISTA [16]—for problem (1) with our proposed IRPN algorithm by applying them to  $\ell_1$ -regularized logistic regression with high-dimensional data sets. Experiment results demonstrate the superiority of IRPN over both **newGLMNET** and the adaptively restarted FISTA. To explore the effect of the parameter  $\rho$  on the behaviour of IRPN, we conduct another experiment by running IRPN on the same problem with different values of  $\rho \in [0, 1]$ . Numerical results corroborate with our theory that the asymptotic convergence rate of IRPN depends on the choice of  $\rho$  and both linear and superlinear rates are observed by tuning  $\rho$ .

The paper is organized as follows. In Section 2, we introduce the assumptions needed through-

out and the said EB condition for (1). In Section 3, we specify the algorithm IRPN, including the regularized Hessian and inexactness conditions for inner minimization. In Sections 4, we establish the global convergence and the local convergence rate of IRPN. Numerical studies are provided in Section 5. Finally, we draw a brief conclusion in Section 6. To increase the readability, all the proofs of the technical results in this paper are presented in the Appendix.

*Notation.* We denote the optimal value of and the set of optimal solutions to problem (1) as  $F^*$  and  $\mathcal{X}$ , respectively, and let  $\|\cdot\|$  be the Euclidean norm. Given a set  $C \subseteq \mathbb{R}^n$ , we denote  $\text{dist}(x, C)$  as the distance from  $x$  to  $C$ ; i.e.,  $\text{dist}(x, C) = \inf\{\|x - u\| \mid u \in C\}$ . For a closed convex function  $h$ , we denote by  $\partial h$  the subdifferential of  $h$ . Furthermore, if  $h$  is smooth, we let  $\nabla h$  and  $\nabla^2 h$  be the gradient and Hessian of  $h$ , respectively.

## 2 Preliminaries

Throughout, we consider problem (1), where the function  $f$  is twice continuously differentiable on an open subset of  $\mathbb{R}^n$  containing  $\text{dom}(g)$ . To avoid ill-posed problems, we assume that the optimal value  $F^* > -\infty$  and the set of optimal solutions  $\mathcal{X}$  is non-empty. Furthermore, we make the following assumptions on the derivatives of  $f$ .

**Assumption 1.** (a). *The gradient of  $f$  is Lipschitz continuous on an open set  $\mathcal{U}$  containing  $\text{dom}(g)$ ; i.e., there exist constants  $L_1 > 0$  such that*

$$\|\nabla f(y) - \nabla f(z)\| \leq L_1 \|y - z\| \quad \forall y, z \in \mathcal{U}. \quad (3)$$

(b). *The Hessian of  $f$  is Lipschitz continuous on an open set  $\mathcal{U}$  containing  $\text{dom}(g)$ ; i.e., there exist constants  $L_2 > 0$  such that*

$$\|\nabla^2 f(y) - \nabla^2 f(z)\| \leq L_2 \|y - z\| \quad \forall y, z \in \mathcal{U}. \quad (4)$$

The above assumptions are standard in the analysis of Newton-type methods. As we will see in Sections 4 and 5, Assumption 1(a) is crucial to the analysis of global convergence and Assumption 1(b) is additionally utilized in analyzing the local convergence rate. A direct consequence of Assumption 1 is that the operator norm of the Hessian  $\nabla^2 f$  is bounded.

**Lemma 1.** *Under Assumption 1, for any  $x \in \text{dom}(g)$ , it follows that  $\lambda_{\max}(\nabla^2 f(x)) \leq L_1$ .*

### 2.1 Optimality and Residual Functions

We are now ready to explore the optimality conditions of (1) and introduce the announced error bound condition. Given a convex function  $h$ , the so-called proximal operator of  $h$  is defined as follows:

$$\text{prox}_h(v) := \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \|u - v\|^2 + h(u). \quad (5)$$

The proximal operator is the building block of many first-order methods for solving (1), such as proximal gradient method and its accelerated versions, and (block) coordinate gradient descent methods and can be viewed as a generalization of projection operator. Indeed, when the function  $h$  is the indicator function of a closed convex set  $C$ , the proximal operator of  $h$  reduces to the

projection operator onto  $C$ . Moreover, the proximal operators of many nonsmooth functions, such as  $\ell_1$ -norm, group-Lasso regularizer, elastic net regularizer and nuclear norm, have closed-form representations. For example, let  $h$  be the scaled  $\ell_1$ -norm function  $\tau\|x\|_1$ . Then the proximal operator of  $h$  is the *soft-thresholding operator*:

$$\text{prox}_h(v) = \text{sign}(v) \odot \max\{|x| - \tau, 0\}$$

where  $\text{sign}$ ,  $\max$ ,  $|\cdot|$  are entrywise operations and  $\odot$  operates the entrywise product. We now recall some of the useful properties of the proximal operator that will be used later. One well-known property of the proximal operator is that it characterizes the optimal solutions of (1) as its fixed points. For its proof, we refer readers to Lemma 2.4 [4].

**Lemma 2.** *A vector  $x \in \text{dom}(g)$  is an optimal solution of (1) if and only if for any  $\tau > 0$ , the following fixed-point equation holds:*

$$x = \text{prox}_{\tau g}(x - \tau \nabla f(x)). \quad (6)$$

Let  $R : \text{dom}(g) \rightarrow \mathbb{R}^n$  be the map given by

$$R(x) := x - \text{prox}_g(x - \nabla f(x)). \quad (7)$$

Lemma 2 suggests that  $r(x) = \|R(x)\|$  is a residual function for problem (1), in the sense that  $r(x) \geq 0$  for all  $x \in \text{dom}(g)$  and  $r(x) = 0$  if and only if  $x \in \mathcal{X}$ . In addition, the following proposition shows that both  $R(x)$  and  $r(x)$  are Lipschitz continuous on  $\text{dom}(g)$  if Assumption 1(a) is satisfied.

**Proposition 1.** *Suppose that Assumption 1(a) is satisfied. Then, for any  $y, z \in \text{dom}(g)$ , we have*

$$|r(y) - r(z)| \leq \|R(y) - R(z)\| \leq (L_1 + 2)\|y - z\|.$$

### 2.1.1 Inner Minimization

Recall that at the  $k$ -th iteration of an SQA method, the goal is to find the minimizer of (2). By letting  $f_k(x)$  be the sum of the first three terms, the inner minimization problem reads

$$\min_{x \in \mathbb{R}^n} q_k(x) := f_k(x) + g(x), \quad (8)$$

which is also a convex minimization of the form (1) with the smooth function being quadratic. Therefore, both the optimality condition and the residual function studied above for problem (1) can be adapted to the inner minimization (8). The following corollary is immediately implied by Lemma 2 after noting the fact that  $\nabla f_k(x) = \nabla f(x_k) + H_k(x - x_k)$ .

**Corollary 1.** *A vector  $x \in \text{dom}(g)$  is a minimizer of (8) if and only if for any  $\tau > 0$ , the following fixed-point equation holds*

$$x = \text{prox}_{\tau g}(x - \tau \nabla f_k(x)) = \text{prox}_{\tau g}((I - \tau H_k)x - \tau(\nabla f(x_k) - H_k x_k)).$$

Similar to the map  $R$  for problem (1), let  $R_k : \text{dom}(g) \rightarrow \mathbb{R}^n$  be the map given by

$$R_k(x) := x - \text{prox}_g(x - \nabla f_k(x)) = x - \text{prox}_g((I - H_k)x - (\nabla f(x_k) - H_k x_k)). \quad (9)$$

Parallel to the role of  $r(x)$  for the outer problem (1),  $r_k(x) = \|R_k(x)\|$  is a natural residual function for the inner minimization (8). Moreover, following the lines of the proof of Proposition 1, we can easily show that both the  $R_k$  and  $r_k$  are Lipschitz continuous

**Corollary 2.** *For any  $y, z \in \mathbb{R}^n$ , we have<sup>2</sup>*

$$|r_k(y) - r_k(z)| \leq \|R_k(y) - R_k(z)\| \leq (\lambda_{\max}(H_k) + 2)\|y - z\|.$$

## 2.2 Error Bound Condition

A prevailing assumption in existing approaches to analyzing the global convergence and local convergence rates of SQA methods for solving problem (1) is that the smooth function  $f$  is strongly convex [10, 3]. However, such an assumption is difficult to verify and is invalid in many applications (see the discussion in Section 1). In this paper, instead of assuming strong convexity, our analysis is based on the following error bound (EB) condition.

**Assumption 2. (EB condition)** *Let  $r(x)$  be the residual function defined in Section 2.1. For any  $\zeta \geq F^*$ , there exist scalars  $\kappa > 0$  and  $\epsilon > 0$  such that*

$$\text{dist}(x, \mathcal{X}) \leq \kappa \cdot r(x) \quad \text{whenever } F(x) \leq \zeta, \quad r(x) \leq \epsilon. \quad (10)$$

Error bounds have long been an important topic and permeate in all aspects of mathematical programming [19, 6]. The specific local error bound (10) for problem (1) is initially studied by Pang [18] for strongly convex smooth function  $f$ , and Luo and Tseng [12, 13, 14] for structured  $f$  and polyhedral function  $g$ . Recently, the results on error bound (10) is extended to non-polyhedral  $g$  [24] and even problems with non-polyhedral optimal solution set [30]. We now briefly summarize the existing results on the validity error bounds in the following theorem. For a survey along this line of research, we refer the readers to the recent manuscript of Zhou and So [30].

**Theorem 1.** *For problem (1), Assumption 2 (EB condition) holds in any of the following scenarios:*

- (S1). ([18, Theorem 3.1])  $f$  is strongly convex,  $\nabla f$  is Lipschitz continuous, and  $g$  is any closed convex function.
- (S2). ([13, Theorem 2.1])  $f$  takes the form  $f(x) = h(Ax) + \langle c, x \rangle$ , where  $A \in \mathbb{R}^{m \times n}$ ,  $c \in \mathbb{R}^n$  are arbitrarily given and  $h : \mathbb{R}^m \rightarrow (-\infty, +\infty)$  is a continuously differentiable function with the property that on any compact subset  $C$  of  $\mathbb{R}^n$ ,  $h$  is strongly convex and  $\nabla h$  is Lipschitz continuous on  $C$ , and  $g$  is of polyhedral epigraph.
- (S3). ([24, Theorem 2])  $f$  takes the form  $f(x) = h(Ax)$ , where  $A \in \mathbb{R}^{m \times n}$  and  $h : \mathbb{R}^m \rightarrow (-\infty, +\infty)$  are as in scenario (S2), and  $g$  is the grouped LASSO regularizer; i.e.,  $g(x) = \sum_{J \in \mathcal{J}} \omega_J \|x_J\|_2$ , where  $\mathcal{J}$  is a non-overlapping partition of the index set  $\{1, \dots, n\}$ ,  $x_J \in \mathbb{R}^{|J|}$  is the sub-vector obtained by restricting  $x \in \mathbb{R}^n$  to the entries in  $J \in \mathcal{J}$ , and  $\omega_J \geq 0$  is a given parameter.

---

<sup>2</sup>Compared with Proposition 1, Assumption 1(a) is not required for Corollary 2.



(S4). ([30, Proposition 12])  $f$  takes the form  $f(X) = h(\mathcal{A}(X))$  for all  $X \in \mathbb{R}^{n \times p}$ , where  $\mathcal{A} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^m$  is a linear mapping,  $h : \mathbb{R}^m \rightarrow (-\infty, +\infty)$  is as in scenario (S2),  $g$  is the nuclear norm regularizer, i.e.,  $g(X)$  equals the sum of singular values of  $X$ , and there exists an  $X^* \in \mathcal{X}$  such that the following strict complementary-type condition holds:

$$\mathbf{0} \in \nabla f(X^*) + \text{ri}(\partial g(X^*)),$$

where  $\text{ri}$  denotes the relative interior.

Note that in scenarios (S2)-(S4), the assumptions on  $h$  are the same and can readily be shown to be satisfied by  $h(y) = \frac{1}{2}\|y - b\|^2$ , which corresponds to least-squares regression, and  $h(y) = \sum_{i=1}^m \log(1 + e^{-b_i y_i})$  with  $b \in \{-1, 1\}^m$ , which corresponds to logistic regression. The assumptions are also satisfied by the loss functions that arise in the maximum likelihood estimation (MLE) for conditional random field problems [29] and MLE under Poisson noise [21]. It follows from Theorem 1 that many problems of the form (1), including  $\ell_1$ -regularized logistic regression, satisfy the EB condition, even when they are not strongly convex.

In Section 4, the EB condition (10) is harnessed to establish the local superlinear convergence rate of our inexact regularized proximal Newton (IRPN) method, thus further demonstrating its versatility in convergence analysis. Prior to this work, the EB condition (10) is used to prove the local linear convergence of a number of first-order methods for solving problem (1), such as the projected gradient method, proximal gradient method, coordinate minimization, and coordinate gradient descent method; see [14, 24, 25] and the references therein. Besides first-order methods, such EB condition is also used to prove the superlinear convergence of primal-dual interior-point path following methods [23]. However, these methods are quite different in nature from the SQA-type methods considered in this paper.

Besides the EB condition (10), there are other regularity conditions that are utilized for establishing the superlinear convergence rate of SQA methods. Yen et al. [27] and Zhong et al. [29] introduced the *constant nullspace strong convexity* (CNSC) condition for a smooth function. They showed that, under the CNSC condition on  $f$  and some other regularity conditions on the nonsmooth function  $g$ , the proximal Newton method [10] converges quadratically and proximal quasi-Newton method [10] converges superlinearly. However, all the instances satisfying their conditions are subsumed by the scenarios in Theorem 1. Moreover, both the inexactness of the inner minimization and the global convergence are left unaddressed. Dontchev and Rockafellar [5] developed a framework of solving generalized equations. When specialized to the optimality condition  $0 \in \nabla f(x) + \partial g(x)$  of problem (1), their framework coincides with the family of SQA methods and the corresponding convergence results require the set-valued mapping  $\nabla f + \partial g$  to be either *metrically regular* or *strongly metrically sub-regular* [5]. Unfortunately, both of these two regularity conditions are provably more restrictive than the EB condition (10) and are hardly satisfied by any instances of (1) of interest. For example, consider the following two-dimensional  $\ell_1$ -norm regularized linear regression:

$$\min_{x_1, x_2} \frac{1}{2}|x_1 + x_2 - 2|^2 + |x_1| + |x_2|.$$

The above problem satisfies the EB condition (10) as it belongs to scenario (S2) in Theorem 1. On the other hand, it can be verified that the set-valued mapping associated with this problem is neither metrically regular nor strongly metrically sub-regular.

### 3 The Inexact Regularized Proximal Newton Method

We now describe in detail the inexact regularized proximal Newton (IRPN) method. The algorithm is input with an initial iterate  $x_0$ , a constant  $\epsilon_0 > 0$  that controls the precision, constants  $\theta \in (0, 1/2), \zeta \in (\theta, 1/2), \eta \in (0, 1)$  that are parameters of inexactness conditions and line search, and constants  $c > 0, \rho \in [0, 1]$  that are used in forming the approximation matrices  $H_k$ 's. As we will see in Section 4, the choice of constant  $\rho$  is crucial for the local convergence rate of IRPN.

At iterate  $x_k$ , we first construct the following quadratic approximation of  $F$  at  $x_k$ ,

$$q_k(x) := f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T H_k(x - x_k) + g(x),$$

where  $H_k = \nabla^2 f(x_k) + \mu_k I$  with  $\mu_k = c \cdot r(x_k)^\rho$ , and  $r(\cdot)$  is the residual function defined in Section 2.1. Since  $f$  is convex and  $\mu_k > 0$ , the matrix  $H_k$  is positive definite for all  $k$ . Hence, the quadratic model  $q_k(x)$  is strictly convex and thus has a unique minimizer. We next find an *approximate* minimizer  $\hat{x}$  of the quadratic model  $q_k(x)$ , such that the following inexactness conditions hold:

$$r_k(\hat{x}) \leq \eta \cdot \min\{r(x_k), r(x_k)^{1+\rho}\} \quad \text{and} \quad q_k(\hat{x}) - q_k(x_k) \leq \zeta(\ell_k(\hat{x}) - \ell_k(x_k)), \quad (11)$$

where  $r_k(\cdot)$  is the residual function for the inner minimization defined in Section 2.1.1 and  $\ell_k$  is the first-order approximation of  $F$  at  $x_k$ :

$$\ell_k(x) := f(x_k) + \nabla f(x_k)^T(x - x_k) + g(x).$$

Since the exact minimizer of  $q_k$  has no closed-form in most cases, an iterative algorithm, such as coordinate minimization, coordinate gradient descent method, and accelerated proximal gradient method, is typically called for finding the approximate minimizer  $\hat{x}$  that satisfies the conditions in (11). After performing a backtracking line search along the direction  $d_k := \hat{x} - x_k$ , we obtain a step size  $\alpha_k > 0$  that guarantees a sufficient decrease of the objective value. The algorithm then steps into the next iterate by setting  $x_{k+1} := x_k + \alpha_k d_k$ . Finally, we terminate the algorithm when  $r(x_k)$  is less than the prescribed constant  $\epsilon_0$ . We summarize the details of IRPN method in Algorithm 1.

The inexactness conditions (11) is similar to those proposed in [3]. However, the analysis in [3] cannot be applied to study the convergence behavior of IRPN. First, the global convergence therein requires  $H_k \succeq \lambda I$  for some  $\lambda > 0$  for all  $k$ . Second, the local convergence rate therein requires  $\nabla^2 f$  to be positive definite at the limiting point  $x^*$  of the sequence  $\{x_k\}_{k \geq 0}$ . Since we do not assume that the smooth function  $f$  is strongly convex, neither of these two assumptions holds in our case.

One advantage of our proposed IRPN lies in its flexibility. Indeed, for SQA methods using regularized Hessian, it is typical to let the regularization parameter  $\mu_k$  be of the same order as the residual function  $r(x_k)$  [11, 20], *i.e.*,  $\mu_k = c \cdot r(x_k)$ . In contrast, we let the order of  $\mu_k$  be adjustable according to the parameter  $\rho$ . Therefore, IRPN is actually a tunable family of algorithms parametrized by  $\rho \in [0, 1]$ . As we will see in Section 4 and Section 5, the parameter  $\rho$  plays a dominant role in the asymptotic convergence rate of IRPN. Even more, IRPN allows one to choose the inner solver. Since each inner problem is a very well-structured minimization problem with objective function being a sum of a strongly convex quadratic function and the



---

**Algorithm 1** Inexact Regularized Proximal Newton (IRPN) Method

---

- 1: **Input:** Starting point  $x_0$  and constants  $\epsilon_0 > 0$ ,  $\theta \in (0, 1/2)$ ,  $\zeta \in (\theta, 1/2)$ ,  $\eta \in (0, 1)$ ,  $c > 0$ ,  $\rho \in [0, 1]$  and  $\beta \in (0, 1)$ .
- 2: **for**  $k = 0, 1, 2, \dots$  **do**
- 3:   Compute the value of residual  $r(x_k)$ .
- 4:   If  $r(x_k) \leq \epsilon_0$ , terminate the algorithm and return  $x_k$ .
- 5:   Form the quadratic model  $q_k(x)$  with  $H_k = \nabla^2 f(x_k) + \mu_k I$  and  $\mu_k = c \cdot r(x_k)^\rho$ .
- 6:   Call an inner solver and find an approximate minimizer  $\hat{x}$  of  $q_k$  such that the following inexactness conditions are satisfied:

$$r_k(\hat{x}) \leq \eta \cdot \min\{r(x_k), r(x_k)^{1+\rho}\} \quad \text{and} \quad q_k(\hat{x}) - q_k(x_k) \leq \zeta(\ell_k(\hat{x}) - \ell_k(x_k)).$$

- 7:   Let the search direction  $d_k := \hat{x} - x_k$ . Find the smallest positive integer  $i$  such that

$$F(x_k) - F(x_k + \beta^i d_k) \geq \theta(\ell_k(x_k) - \ell_k(x_k + \beta^i d_k)). \quad (12)$$

- 8:   Let the step size  $\alpha_k = \beta^i$  and set  $x_{k+1} = x_k + \alpha_k d_k$ .

9: **end for**

---

nonsmooth convex function  $g$ , diligent users can speed up the inner minimization by exploiting the structure of  $\nabla^2 f$  and  $g$  to design special inner solvers.

Before we end this section, we show that the IRPN method is well defined. We first present the following lemma, which ensures that at each iteration, the inexactness conditions (11) are satisfied as long as  $\hat{x}$  is accurate enough to the exact solution of the inner minimization. The proof is identical to Lemma 4.5 in [3], but we give a proof in the Appendix for completeness.

**Lemma 3.** *The inexactness conditions (11) are satisfied for any vectors in  $\mathbb{R}^n$  that are sufficiently close to the exact minimizer  $\bar{x}$  of  $q_k(x)$ .*

We next show that the backtracking line search is well defined.

**Lemma 4.** *Suppose that Assumption 1(a) is satisfied. Then, for any iteration  $k$  of Algorithm 1, there exists a positive integer  $i$  such that the descent condition in (12) is satisfied. Moreover, the step size  $\alpha_k$  obtained from the line search strategy satisfies*

$$\alpha_k \geq \frac{\beta(1-\theta)\mu_k}{(1-\zeta)L_1}.$$

Combining Lemma 3 and Lemma 4, we conclude that the method IRPN is well defined.

## 4 Convergence Analysis

Throughout this section, we denote  $\{x_k\}_{k \geq 0}$  as the sequence of iterates generated by Algorithm 1. We first present the a theorem showing the global convergence of Algorithm 1.

**Theorem 2** (Global Convergence). *Suppose that Assumption 1(a) holds. Then, we have*

$$\lim_{k \rightarrow \infty} r(x_k) = 0. \quad (13)$$

In other words, each accumulation point of the sequence  $\{x_k\}_{k \geq 0}$  is an optimal solution of (1).

Having established the global convergence of the IRPN method, we now study its asymptotic convergence rate. We show that remarkably, the IRPN method can attain superlinear and quadratic rates of convergence even without strong convexity (see Theorem 3). Since we do not assume  $f$  to be strongly convex, techniques based on the facts that the Hessian of  $f$  is positive definite at an optimal  $x^*$  or the optimal solution is unique collapse. The key to the proof of Theorem 3 is the EB condition (10). Our technique is novel and we believe that it will find further applications in the convergence rate analysis of other second-order methods in the absence of strong convexity.

**Theorem 3** (Local Convergence Rate). *Suppose that Assumptions 1 and 2 hold. Then, for sufficiently large  $k$ , we have*

- (i)  $d(x_{k+1}, \mathcal{X}) \leq \gamma d(x_k, \mathcal{X})$  for some  $\gamma \in (0, 1)$ , if we take  $\rho = 0$ ,  $c \leq \frac{\kappa}{4}$ , and  $\eta \leq \frac{1}{2(L_1+3)^2}$ ;
- (ii)  $d(x_{k+1}, \mathcal{X}) \leq O(d(x_k, \mathcal{X})^{1+\rho})$  if we take  $\rho \in (0, 1)$ ;
- (iii)  $d(x_{k+1}, \mathcal{X}) \leq O(d(x_k, \mathcal{X})^2)$  if we take  $\rho = 1$ ,  $c \geq \kappa L_2$ , and  $\eta \leq \frac{\kappa^2 L_2}{2(L_1+1)}$ .

We remark that though Theorem 3 shows that a larger  $\rho$  leads to faster convergence rate, it does not suggest that a larger  $\rho$  leads to a faster algorithm with respect to time complexity. The reason is that a larger  $\rho$  results in stricter inexactness conditions and thus the time consumed by each iteration increases. Hence, despite the faster convergence rate, the best choice of  $\rho \in [0, 1]$  depends on particular problems and the solver used for inner minimization. Nonetheless, Theorem 3 provides a complete characterization of the convergence rate of IRPN in terms of  $\rho$  and the flexibility of IRPN is manifested through this parameter.

## 5 Numerical Experiments

We now explore the numerical performance of the proposed IRPN algorithm. Specifically, we apply it to solve  $\ell_1$ -regularized logistic regression, which is widely used for linear classification tasks and also a standard benchmark problem for testing the efficiency of an algorithm for solving (1). The optimization problem of  $\ell_1$ -regularized logistic regression takes the following form:

$$\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-b_i \cdot a_i^T x)) + \mu \|x\|_1, \quad (14)$$

where  $(a_i, b_i) \in \mathbb{R}^n \times \{-1, 1\}$ ,  $i = 1, \dots, n$  are data points. We employ the data set RCV1, which contains 20242 samples and 47236 features and is downloaded from LIBSVM datasets repository. Since the number of features is larger than the number of data points, the objective function  $F(x)$  in (14) associated with this classification task is not strongly convex. However, due to Theorem 1, the error bound condition (10) holds. Hence, both Assumption 1 and Assumption 2 hold for problem (14) and our convergence analysis of IRPN applies when solving problem (14).

We compare our algorithm IRPN with two venerable methods for solving (14). The first one is the newGLMNET algorithm, which is also a SQA method and is the workhorse of the well-known LIBLINEAR package [28]. The second one for comparison is the fast iterative shrinkage algorithm

with adaptive restarting scheme [16] (**adaFISTA**), which is a first-order method. Although both IRPN and **newGLMNET** are instantiations of SQA methods, there are two main differences: (i) **newGLMNET** regularizes the Hessian by adding a constant multiple of identity matrix while IRPN employs an adaptive regularization (see Step 5 of Algorithm 1); (ii) **newGLMNET** uses heuristic stopping schemes for inner minimization while IRPN employs adaptive inexactness conditions (11) that leads to provable convergence guarantees. To highlight these two differences, we deactivate other heuristic tricks in the implementation that can be applied to general SQA methods for acceleration, such as shrinking strategy. Moreover, for both of these two methods, randomized coordinate descent algorithm are called for solving the inner minimization.

Tol.	Algorithm	adaFISTA	newGLMNET	IRPN $\rho = 0$	IRPN $\rho = 0.5$	IRPN $\rho = 1$
$10^{-4}$	Outer iterations	264	21	14	9	<b>8</b>
	Inner iterations	–	52	<b>28</b>	51	54
	Training time (sec)	4.19	4.10	<b>3.56</b>	6.10	4.53
$10^{-6}$	Outer iterations	450	28	20	10	<b>9</b>
	Inner iterations	–	85	<b>70</b>	92	123
	Training time (sec)	7.14	7.65	<b>6.02</b>	7.90	11.34
$10^{-8}$	Outer iterations	1978	38	23	11	<b>9</b>
	Inner iterations	–	116	<b>90</b>	151	146
	Training time (sec)	31.52	10.56	<b>9.32</b>	13.44	12.10

Table 1: Comparison of number of iterations and time consumed. Algorithms are terminated when the residual  $r(x)$  is less than the tolerances  $10^{-4}$ ,  $10^{-6}$  and  $10^{-8}$ .

We see from Table 1 that IRPN with  $\rho = 0$  is the best among the compared algorithms in the sense that it requires minimum training time for reaching any of the given tolerances. Note that when  $\rho = 0$ , IRPN differs with **newGLMNET** only on the stopping conditions for inner minimization. Hence, Table 1 suggests that the adaptive inexactness conditions (11) is not only theoretically sound but also practically better than the heuristic stopping scheme employed by **newGLMNET**.

We then study the convergence rate of the sequence  $\{x_k\}_{k \geq 0}$  generated by the IRPN method. Thanks to the error bound condition (10), the speed that  $x_k$  converges to the optimal set  $\mathcal{X}$  is asymptotically of the order  $r(x_k)$ . Hence, it suffices to investigate the rate of  $r(x_k)$  converging to 0. It can be observed from Figure 1(a) that  $r(x_k)$  converges linearly if  $\rho = 0$  and superlinearly if  $\rho = 0.5$  or  $\rho = 1$ . This corroborates with our results in Theorem 3. Nonetheless, a faster convergence rate of  $r(x_k)$  does not tell the whole story since a larger  $\rho$  will lead to stricter inexactness conditions (11) and hence more iterations for each inner minimization are needed. We refer to Figure 1(b), which plots the decreases of (logarithms of)  $r(x_k)$  against training time, for the overall performance. The figure shows that the training time complexity of these three values of  $\rho$  are almost the same.

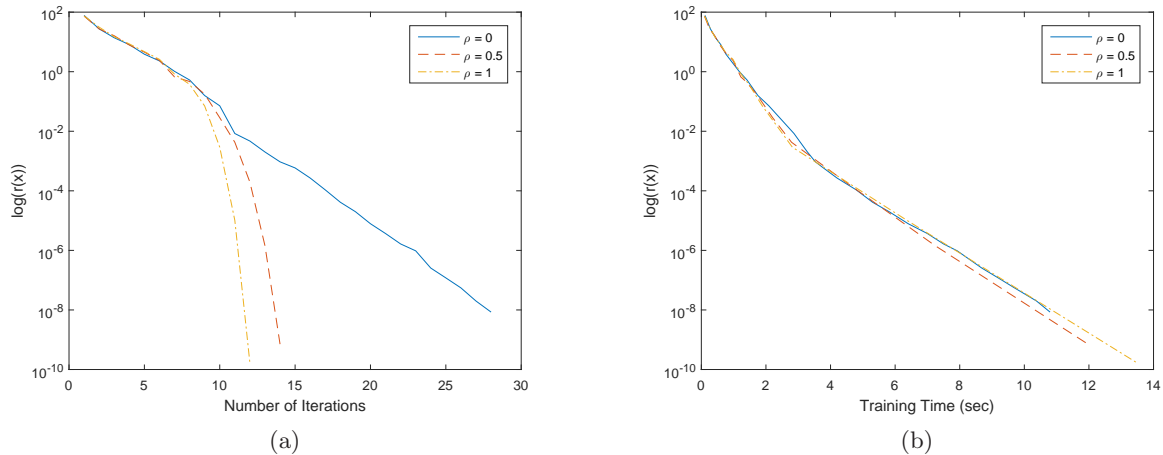


Figure 1: Convergence performance of the IRPN algorithm.

## 6 Conclusions

In this paper, we propose an *inexact regularized proximal Newton* (IRPN) method, which belongs to the class of successive quadratic approximation methods, for minimizing a sum of smooth and non-smooth convex functions. Based on a certain error bound condition, we prove that the IRPN method globally converges in the sense that the distance of the iterates generated by IRPN converges to zero and the asymptotic rate of convergence is superlinear, even when the objective function is not strongly convex. Although this error bound played a fundamental role in establishing linear convergence of various first-order methods, to the best of our knowledge, it is the first time that such condition is utilized for proving superlinear convergence of SQA methods. We compare our proposed IRPN with two popular and efficient algorithms—the newGLMNET [28] and adaptively restarted FISTA [16]—by applying them to the  $\ell_1$ -regularized logistic regression problem. Experiment results indicate that IRPN achieves the same accuracies in shorter training time and less number of iterations than the two other algorithms. This shows that the IRPN method does not only come with great flexibility and a complete and promising theory, but also superior numerical performance.

# Appendix

## A Technical Proofs

### A.1 Proof of Lemma 1

*Proof of Lemma 1.* Suppose  $x$  is an arbitrary vector in  $\text{dom}(g)$ . By Assumption 1,  $f$  is twice continuously differentiable at  $x$ . Hence, by the definition of  $\nabla^2 f$ , we have for any vector  $v \in \mathbb{R}^n$ ,

$$\nabla^2 f(x)v = \lim_{h \rightarrow 0} \frac{\nabla f(x + hv) - \nabla f(x)}{h}.$$

By taking the vector norm on both sides, we have

$$\|\nabla^2 f(x)v\| = \lim_{h \rightarrow 0} \frac{\|\nabla f(x+hv) - \nabla f(x)\|}{h} \leq \lim_{h \rightarrow 0} \frac{L_1 h \|v\|}{h} = L_1 \|v\|,$$

where the inequality is due to Assumption 1(a). Hence, by the definition of  $\lambda_{\max}$ , we have  $\lambda_{\max}(\nabla^2 f(x)) \leq L_1$ .  $\square$

## A.2 Proof of Proposition 1

*Proof of Proposition 1.* The first inequality is a direct consequence of the triangle inequality. We now prove the second one. It follows from the definition of  $R$  that

$$\begin{aligned} \|R(y) - R(z)\| &= \|y - \text{prox}_g(y - \nabla f(y)) - z + \text{prox}_g(z - \nabla f(z))\| \\ &\leq \|y - z\| + \|\text{prox}_g(y - \nabla f(y)) - \text{prox}_g(z - \nabla f(z))\| \\ &\leq 2\|y - z\| + \|\nabla f(y) - \nabla f(z)\| \\ &\leq (L_1 + 2)\|y - z\|, \end{aligned}$$

where the second inequality follows from the firm non-expansiveness of the proximal operator (see Proposition 3.5 of [4]) and the last one is by Assumption 1(a).  $\square$

## A.3 Well-definedness of IRPN

*Proof of Lemma 3.* Suppose we are at iteration  $k$  with the iterates  $x_k$ . In addition, suppose  $x_k$  is not optimal for problem (1) because otherwise the algorithm has terminated. The first condition in (11) is readily verified by noting that  $r_k(\bar{x}) = 0$  and  $r_k(\cdot)$  is continuous. It remains to show that the second condition in (11) will eventually be satisfied. Since  $q_k(x) = \ell_k(x) + \frac{1}{2}(x - x_k)^T H_k(x - x_k)$  and  $\bar{x}$  is the minimizer of  $q_k$ , we have  $H_k(x_k - \bar{x}) \in \partial \ell_k(\bar{x})$ . This leads to the property that  $x_k \neq \bar{x}$ , because otherwise we have  $\mathbf{0} \in \partial \ell_k(x_k)$ , which implies that  $x_k$  is an optimal solution of (1). Moreover, by the convexity of  $\ell_k$ , we have

$$\ell_k(x_k) \geq \ell_k(\bar{x}) + (\bar{x} - x_k)^T H_k(\bar{x} - x_k). \quad (15)$$

Since  $H_k = \nabla^2 f(x_k) + \mu_k I$  for some  $\mu_k > 0$ ,  $H_k$  is positive definite. This, together with the fact that  $x_k \neq \bar{x}$ , implies that  $\ell_k(x_k) > \ell_k(\bar{x})$ . Furthermore, we have

$$\begin{aligned} q_k(\bar{x}) - q_k(x_k) &= \ell_k(\bar{x}) + \frac{1}{2}(\bar{x} - x_k)^T H_k(\bar{x} - x_k) - \ell_k(x_k) \\ &\leq \frac{1}{2}(\ell_k(\bar{x}) - \ell_k(x_k)) \\ &< \zeta(\ell_k(\bar{x}) - \ell_k(x_k)) \end{aligned}$$

where the first inequality is due to (15) and the last one is due to the facts  $\zeta \in (\theta, 1/2)$  and  $\ell_k(x_k) > \ell_k(\bar{x})$ . Therefore, due to the continuity of  $q_k$  and  $\ell_k$ , the second condition in (11) is satisfied as long as  $x$  is close enough to  $\bar{x}$ .  $\square$

*Proof of Lemma 4.* By definition, we have

$$q_k(\hat{x}) - q_k(x_k) = \ell_k(\hat{x}) - \ell_k(x_k) + \frac{1}{2}(\hat{x} - x_k)^T H_k(\hat{x} - x_k).$$

Since the approximate minimizer  $\hat{x}$  satisfies the inexactness conditions (11), the above implies

$$\ell_k(x_k) - \ell_k(\hat{x}) \geq \frac{1}{2(1-\zeta)}(\hat{x} - x_k)^T H_k(\hat{x} - x_k) \geq \frac{\mu_k}{2(1-\zeta)}\|\hat{x} - x_k\|^2. \quad (16)$$

Due to the convexity of  $\ell_k(x)$ , for any constant  $\alpha \in [0, 1]$ , we have

$$\ell_k(x_k) - \ell_k(x_k + \alpha d_k) \geq \alpha(\ell_k(x_k) - \ell_k(x_k + d_k)).$$

Combining the above two inequalities and noting that  $\hat{x} = x_k + d_k$ , we obtain

$$\ell_k(x_k) - \ell_k(x_k + \alpha d_k) \geq \frac{\alpha\mu_k}{2(1-\zeta)}\|d_k\|^2. \quad (17)$$

Moreover, since  $\nabla f$  is Lipschitz continuous by Assumption 1(a), it then follows that

$$f(x_k + \alpha d_k) - f(x_k) \leq \alpha \nabla f(x_k)^T d_k + \frac{\alpha^2 L_1}{2}\|d_k\|^2, \quad (18)$$

which leads to, by the definition of  $\ell_k$ ,

$$F(x_k) - F(x_k + \alpha d_k) \geq \ell_k(x_k) - \ell_k(x_k + \alpha d_k) - \frac{\alpha^2 L_1}{2}\|d_k\|^2.$$

It then follows from the above inequality and (17) that

$$\begin{aligned} & F(x_k) - F(x_k + \alpha d_k) - \theta(\ell_k(x_k) - \ell_k(x_k + \alpha d_k)) \\ & \geq (1 - \theta)(\ell_k(x_k) - \ell_k(x_k + \alpha d_k)) - \frac{\alpha^2 L_1}{2}\|d_k\|^2 \\ & \geq \frac{(1 - \theta)\alpha\mu_k}{2(1 - \zeta)}\|d_k\|^2 - \frac{\alpha^2 L_1}{2}\|d_k\|^2 \\ & = \frac{\alpha}{2} \left( \frac{1 - \theta}{1 - \zeta} \mu_k - \alpha L_1 \right) \|d_k\|^2. \end{aligned}$$

Hence, as long as  $\alpha$  satisfies  $\alpha < (1 - \theta)\mu_k/(1 - \zeta)L_1$ , the descent condition (12) is satisfied. Since the backtracking line search multiplies the step-length by  $\beta \in (0, 1)$  after each trial, then the line search strategy will output an  $\alpha_k$  that satisfies  $\alpha_k \geq \beta(1 - \theta)\mu_k/(1 - \zeta)L_1$ .  $\square$

#### A.4 Global Convergence

*Proof of Theorem 2.* From the inequality (17) in Section A.3, we obtain

$$\ell_k(x_k) - \ell_k(x_k + \alpha d_k) \geq \frac{\alpha\mu_k}{2(1-\zeta)}\|d_k\|^2 \geq 0. \quad (19)$$

Hence, due to the descent condition (12) and the assumption that  $\inf F > -\infty$ , we have

$$\lim_{k \rightarrow \infty} \ell_k(x_k) - \ell_k(x_k + \alpha_k d_k) = 0.$$



Again, using (17), the above implies  $\lim_{k \rightarrow \infty} \alpha_k \mu_k \|d_k\|^2 = 0$ . Moreover, recall that the inexactness condition (11) implies  $r_k(\hat{x}) \leq \eta \cdot r(x_k)$ . It then follows that

$$(1 - \eta)r(x_k) \leq r(x_k) - r_k(\hat{x}) = r_k(x_k) - r_k(\hat{x}) \leq (\lambda_{\max}(H_k) + 2)\|d_k\| \leq (L_1 + \mu_k + 2)\|d_k\|,$$

where the first equality follows from  $r_k(x_k) = r(x_k)$ , the second inequality is by Corollary 2 and the last inequality is due to Lemma 1. Then, combining Lemma 4,  $\lim_{k \rightarrow \infty} \alpha_k \mu_k \|d_k\|^2 = 0$  and  $(1 - \eta)r(x_k) \leq (L_1 + \mu_k + 2)\|d_k\|$ , we have

$$\lim_{k \rightarrow \infty} \frac{\mu_k^2}{(\mu_k + L_1 + 2)^2} r(x_k)^2 = 0.$$

This, together with the fact that  $\mu_k = c \cdot r(x_k)^\rho$ , implies that  $\lim_{k \rightarrow \infty} r(x_k) = 0$ .  $\square$

## A.5 Local Convergence

To increase comprehensibility, we divide the proof of Theorem 3 into a series of lemmas. Since these lemmas are all proved under the conditions of Theorem 3, we suppress the required assumptions in their statements.

**Lemma 5.** *Let  $\hat{x}_{ex}$  be the exact solution of the  $k$ -th subproblem, i.e.  $r_k(\hat{x}_{ex}) = 0$ , and  $\bar{x}_k$  be the projection of  $x_k$  onto the optimal set  $\mathcal{X}$ . Then*

$$\|\hat{x}_{ex} - \bar{x}_k\| \leq \frac{1}{\mu_k} \|\nabla f(\bar{x}_k) - \nabla f(x_k) - H_k(\bar{x}_k - x_k)\|. \quad (20)$$

*Proof.* If  $\hat{x}_{ex} = \bar{x}_k$ , the inequality holds trivially. Assuming  $\hat{x}_{ex} \neq \bar{x}_k$ , it follows from definitions that  $0 \in \nabla f(\bar{x}_k) + \partial g(\bar{x}_k)$  and  $0 \in \nabla f(x_k) + H_k(\hat{x}_{ex} - x_k) + \partial g(\hat{x}_{ex})$ . Since  $g$  is convex, the subdifferential is monotone. Hence we have

$$\begin{aligned} 0 &\leq \langle \nabla f(x_k) - \nabla f(\bar{x}_k) + H_k(\hat{x}_{ex} - x_k), \bar{x}_k - \hat{x}_{ex} \rangle \\ &= \langle \nabla f(x_k) - \nabla f(\bar{x}_k) - H_k(x_k - \bar{x}_k), \bar{x}_k - \hat{x}_{ex} \rangle - \langle H_k(\bar{x}_k - \hat{x}_{ex}), \bar{x}_k - \hat{x}_{ex} \rangle \\ &\leq \|\nabla f(x_k) - \nabla f(\bar{x}_k) - H_k(x_k - \bar{x}_k)\| \|\bar{x}_k - \hat{x}_{ex}\| - \mu_k \|\bar{x}_k - \hat{x}_{ex}\|^2, \end{aligned} \quad (21)$$

which yields the desired inequality.  $\square$

We remark here that we did not use any special structure of  $H_k$  in the proof of Lemma 5 but the assumption that  $H_k$  is positive definite. Therefore this lemma also applies to other positive definite approximate Hessian  $H_k$  with  $\mu_k$  replaced by the minimum eigenvalue  $\lambda_{\min}(H_k)$ .

**Lemma 6.** *It holds for all  $k$  that*

$$\|\hat{x}_{ex} - x_k\| \leq \left( \frac{L_2}{2\mu_k} \text{dist}(x_k, \mathcal{X}) + 2 \right) \text{dist}(x_k, \mathcal{X}). \quad (22)$$

*Proof.* Using triangle inequality and the definition of  $\mu_k$ ,

$$\begin{aligned} \|\nabla f(\bar{x}_k) - \nabla f(x_k) - H_k(\bar{x}_k - x_k)\| &\leq \|\nabla f(\bar{x}_k) - \nabla f(x_k) - \nabla^2 f(x_k)(\bar{x}_k - x_k)\| + \mu_k \|\bar{x}_k - x_k\| \\ &\leq \frac{L_2}{2} \|\bar{x}_k - x_k\|^2 + \mu_k \|\bar{x}_k - x_k\|, \end{aligned} \quad (23)$$

where the second inequality follows from Assumption 1(b). Then, by Lemma 5 and inequality (23),

$$\begin{aligned}
\|\hat{x}_{ex} - x_k\| &\leq \|\hat{x}_{ex} - \bar{x}_k\| + \text{dist}(x_k, \mathcal{X}) \\
&\leq \frac{1}{\mu_k} \|\nabla f(\bar{x}_k) - \nabla f(x_k) - H_k(\bar{x}_k - x_k)\| + \text{dist}(x_k, \mathcal{X}) \\
&\leq \frac{L_2}{2\mu_k} \|\bar{x}_k - x_k\|^2 + \|\bar{x}_k - x_k\| + \text{dist}(x_k, \mathcal{X}) \\
&= \left( \frac{L_2}{2\mu_k} \text{dist}(x_k, \mathcal{X}) + 2 \right) \text{dist}(x_k, \mathcal{X})
\end{aligned} \tag{24}$$

□

**Lemma 7.** *It holds for all  $k$  that*

$$\|\hat{x} - \hat{x}_{ex}\| \leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x_k) + \eta r(x_k)^{1+\rho}. \tag{25}$$

*Proof.* The inequality holds trivially if  $\hat{x} = \hat{x}_{ex}$ . So we assume  $\hat{x} \neq \hat{x}_{ex}$ . Recall that  $R_k(\hat{x}) = \hat{x} - \text{prox}_g(\hat{x} - \nabla f_k(\hat{x}))$ . Then

$$\begin{aligned}
R_k(\hat{x}) - \nabla f_k(\hat{x}) &\in \partial g(\hat{x} - R_k(\hat{x})) \\
R_k(\hat{x}) + \nabla f_k(\hat{x} - R_k(\hat{x})) - \nabla f_k(\hat{x}) &\in \partial q_k(\hat{x} - R_k(\hat{x})) \\
(I - H_k)R_k(\hat{x}) &\in \partial q_k(\hat{x} - R_k(\hat{x}))
\end{aligned} \tag{26}$$

On the other hand, we have  $0 \in \partial q_k(\hat{x}_{ex})$ . Since  $q_k = f_k + g$  is  $\mu_k$ -strongly convex,  $\partial q_k$  is  $\mu_k$ -strongly monotone and thus

$$\langle (I - H_k)R_k(\hat{x}), \hat{x} - R_k(\hat{x}) - \hat{x}_{ex} \rangle \geq \mu_k \|\hat{x} - R_k(\hat{x}) - \hat{x}_{ex}\|^2. \tag{27}$$

Then using Cauchy-Schwarz inequality,

$$\begin{aligned}
\|\hat{x} - R_k(\hat{x}) - \hat{x}_{ex}\| &\leq \frac{1}{\mu_k} \|(I - H_k)R_k(\hat{x})\| \\
&\leq \frac{1}{\mu_k} \|\nabla^2 f(x_k) - (1 - \mu_k)I\|_{op} \cdot r_k(\hat{x}) \\
&\leq \frac{\eta(L_1 + |1 - \mu_k|)}{\mu_k} r(x_k)^{1+\rho} \\
&\leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x_k),
\end{aligned} \tag{28}$$

where the third inequality is due to Assumption 1(c) and the inexactness condition on  $r_k(\hat{x})$ , and the last inequality is due to the setting of  $\mu_k$ . Hence,

$$\begin{aligned}
\|\hat{x} - \hat{x}_{ex}\| &\leq \|\hat{x} - R_k(\hat{x}) - \hat{x}_{ex}\| + \|R_k(\hat{x})\| \\
&\leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x_k) + \eta r(x_k)^{1+\rho}.
\end{aligned} \tag{29}$$

□

We next present a lemma showing that eventually we have unit step-lengths, i.e.  $\alpha_k = 1$ , and the direction  $d_k$  is directly used without doing line search.

**Lemma 8.** Suppose that Assumptions 1 and 2 hold. Then there exists a positive integer  $k_0$  such that  $\alpha_k = 1$  for all  $k \geq k_0$

(i) if  $\rho \in [0, 1)$ , or

(ii) if  $\rho = 1$  and  $c, \eta$  satisfy  $2\eta L_2(L_1 + 2) + \kappa^2 L_2^2 + 2c\kappa L_2 \leq 6c^2$ .

In particular, the inequality in (ii) is satisfied if we take  $c \geq \kappa L_2$  and  $\eta \leq \frac{\kappa^2 L_2}{2(L_1 + 1)}$ .

*Proof.* First note that

$$\begin{aligned}
f(\hat{x}) - f(x_k) &= (\hat{x} - x_k)^T \int_0^1 \nabla f(x_k + t(\hat{x} - x_k)) dt \\
&= (\hat{x} - x_k)^T \int_0^1 [\nabla f(x_k + t(\hat{x} - x_k)) - \nabla f(x_k)] dt + (\hat{x} - x_k)^T \nabla f(x_k) \\
&= (\hat{x} - x_k)^T \int_0^1 t [\nabla^2 f(x_k + st(\hat{x} - x_k)) - \nabla^2 f(x_k)] ds dt (\hat{x} - x_k) \\
&\quad + (\hat{x} - x_k)^T \int_0^1 t \nabla^2 f(x_k) dt (\hat{x} - x_k) + \nabla f(x_k)^T (\hat{x} - x_k) \\
&\leq \frac{1}{2} (\hat{x} - x_k)^T \nabla^2 f(x_k) (\hat{x} - x_k) + \nabla f(x_k)^T (\hat{x} - x_k) + \frac{L_2}{6} \|\hat{x} - x_k\|^3.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&F(\hat{x}) - F(x_k) + q_k(x_k) - q_k(\hat{x}) \\
&= f(\hat{x}) - f(x_k) - \nabla f(x_k)^T (\hat{x} - x_k) - \frac{1}{2} (\hat{x} - x_k)^T \nabla^2 f(x_k) (\hat{x} - x_k) - \frac{\mu_k}{2} \|\hat{x} - x_k\|^2 \\
&\leq \frac{L_2}{6} \|\hat{x} - x_k\|^3 - \frac{\mu_k}{2} \|\hat{x} - x_k\|^2.
\end{aligned} \tag{30}$$

Hence using inequalities (30), (11) and (16), we have

$$\begin{aligned}
F(\hat{x}) - F(x_k) &= (F(\hat{x}) - F(x_k) - q_k(\hat{x}) + q_k(x_k)) + (q_k(\hat{x}) - q_k(x_k)) \\
&\leq \zeta (\ell_k(\hat{x}) - \ell_k(x_k)) + \frac{L_2}{6} \|d_k\|^3 - \frac{\mu_k}{2} \|d_k\|^2 \\
&= \theta (\ell_k(\hat{x}) - \ell_k(x_k)) + (\zeta - \theta) (\ell_k(\hat{x}) - \ell_k(x_k)) + \frac{L_2}{6} \|d_k\|^3 - \frac{\mu_k}{2} \|d_k\|^2 \\
&\leq \theta (\ell_k(\hat{x}) - \ell_k(x_k)) - \left(1 + \frac{\zeta - \theta}{1 - \zeta}\right) \frac{\mu_k}{2} \|d_k\|^2 + \frac{L_2}{6} \|d_k\|^3 \\
&\leq \theta (\ell_k(\hat{x}) - \ell_k(x_k))
\end{aligned} \tag{31}$$

if

$$\|d_k\| = \|\hat{x} - x_k\| \leq \left(1 + \frac{\zeta - \theta}{1 - \zeta}\right) \frac{3cr(x_k)^\rho}{L_2}. \tag{32}$$

By Lemma 6 and Lemma 7, for sufficiently large  $k$ ,

$$\begin{aligned}
\|\hat{x} - x_k\| &\leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x_k) + \eta r(x_k)^{1+\rho} + \left(\frac{L_2}{2cr(x_k)^\rho} \text{dist}(x_k, \mathcal{X}) + 2\right) \text{dist}(x_k, \mathcal{X}) \\
&\leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x_k) + \eta r(x_k)^{1+\rho} + \frac{\kappa^2 L_2}{2c} r(x_k)^{2-\rho} + 2\kappa r(x_k).
\end{aligned} \tag{33}$$

Hence a sufficient condition for inequality (32) is

$$\frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x_k) + \eta r(x_k)^{1+\rho} + \frac{\kappa^2 L_2}{2c} r(x_k)^{2-\rho} + 2\kappa r(x_k) \leq \left(1 + \frac{\zeta - \theta}{1 - \zeta}\right) \frac{3cr(x_k)^\rho}{L_2}. \quad (34)$$

If  $\rho \in [0, 1)$ , then since  $r(x_k) \rightarrow 0$ , (34) holds for sufficiently large  $k$ . Now we consider the case  $\rho = 1$ . Since  $\zeta > \theta$  and  $\mu_k \in (0, 1)$  for sufficiently large  $k$ , a sufficient condition for (34) is

$$\begin{aligned} \frac{\eta(L_1 + 1)}{c} + \frac{\kappa^2 L_2}{2c} + 2\kappa &\leq \frac{3c}{L_2} \\ 2\eta L_2(L_1 + 1) + \kappa^2 L_2^2 + 2c\kappa L_2 &\leq 6c^2, \end{aligned}$$

which in particular holds if we take  $\eta \leq \frac{\kappa^2 L_2}{2(L_1 + 1)}$  and  $c \geq \kappa L_2$ .  $\square$

We finally have enough tools at our disposal to prove Theorem 3.

*Proof of Theorem 3.* From the proof of Lemma 8 and Proposition 1, we have

$$\begin{aligned} \|\hat{x} - x_k\| &\leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x_k) + \eta r(x_k)^{1+\rho} + \left( \frac{L_2}{2cr(x_k)^\rho} \text{dist}(x_k, \mathcal{X}) + 2 \right) \text{dist}(x_k, \mathcal{X}) \\ &\leq \left( \frac{\eta(L_1 + |1 - \mu_k|)(L_1 + 2)}{c} + \eta(L_1 + 2)r(x_k)^\rho + \frac{L_2}{2cr(x_k)^\rho} \text{dist}(x_k, \mathcal{X}) + 2 \right) \text{dist}(x_k, \mathcal{X}) \end{aligned} \quad (35)$$

By the error bound assumption,  $\text{dist}(x, \mathcal{X}) \leq \kappa r(x)$  for sufficiently small  $r(x)$ . Since  $r(x_k) \rightarrow 0$  as  $k \rightarrow \infty$ , we have that for sufficiently large  $k$ ,

$$\|\hat{x} - x_k\| = O(\text{dist}(x_k, \mathcal{X})). \quad (36)$$

Using the same arguments as in Lemma 6, it can be shown that

$$\|\text{prox}_g(\hat{x} - \nabla f(\hat{x})) - \text{prox}_g(\hat{x} - \nabla f_k(\hat{x}))\| \leq \|\nabla f(\hat{x}) - \nabla f(x_k) - H_k(\hat{x} - x_k)\|. \quad (37)$$

Finally, we have that for sufficiently large  $k$ ,

$$\begin{aligned} \text{dist}(\hat{x}, \mathcal{X}) &\leq \kappa r(\hat{x}) \\ &\leq \kappa \|R(\hat{x}) - R_k(\hat{x})\| + \kappa \|R_k(\hat{x})\| \\ &\leq \kappa \|\text{prox}_g(\hat{x} - \nabla f(\hat{x})) - \text{prox}_g(\hat{x} - \nabla f_k(\hat{x}))\| + \kappa \eta r(x_k)^{1+\rho} \\ &\leq \kappa \|\nabla f(\hat{x}) - \nabla f(x_k) - H_k(\hat{x} - x_k)\| + \kappa \eta r(x_k)^{1+\rho} \\ &\leq \frac{\kappa L_2}{2} \|\hat{x} - x_k\|^2 + c\kappa r(x_k)^\rho \|\hat{x} - x_k\| + \kappa \eta r(x_k)^{1+\rho} \\ &\leq \frac{\kappa L_2}{2} \|\hat{x} - x_k\|^2 + c\kappa (L_1 + 2)^\rho \text{dist}(x_k, \mathcal{X})^\rho \|\hat{x} - x_k\| \\ &\quad + \kappa \eta (L_1 + 2)^{1+\rho} \text{dist}(x_k, \mathcal{X})^{1+\rho} \\ &\leq O(\text{dist}(x_k, \mathcal{X})^{1+\rho}), \end{aligned} \quad (38)$$

where first inequality follows from the error bound assumption, the fourth from (37), the sixth from Proposition 1, and the seventh from (35).

If  $\rho = 0$ , inequality (38) does not necessarily imply linear convergence unless the constant in the big-O notation is strictly less than 1. Inspecting the second last line of (38) and the second line of (35), a sufficient condition is

$$\kappa\eta(L_1 + 2) + c\kappa \left( \frac{\eta(L_1 + |1 - c|)(L_1 + 2)}{c} + \eta(L_1 + 2) + 2 \right) < 1, \quad (39)$$

which in particular holds if we take  $\eta \leq \frac{1}{2(L_1+3)^2}$  and  $c \leq \kappa/4$ .  $\square$

## References

- [1] A. Beck and M. Teboulle. A Fast Iterative Shrinkage–Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- [2] S. Becker and J. Fadili. A Quasi–Newton Proximal Splitting Method. In *Advances in Neural Information Processing Systems*, pages 2618–2626, 2012.
- [3] R. H. Byrd, J. Nocedal, and F. Oztoprak. An Inexact Successive Quadratic Approximation Method for L–1 Regularized Optimization. Accepted for publication in *Mathematical Programming, Series B*, 2015.
- [4] P. L. Combettes and V. R. Wajs. Signal Recovery by Proximal Forward–Backward Splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2005.
- [5] A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings*. Springer Monographs in Mathematics. Springer Science+Business Media, LLC, New York, 2009.
- [6] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.
- [7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.
- [8] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, et al. Pathwise Coordinate Optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.
- [9] C.-J. Hsieh, I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik. Sparse Inverse Covariance Matrix Estimation using Quadratic Approximation. In *Advances in Neural Information Processing Systems*, pages 2330–2338, 2011.
- [10] J. D. Lee, Y. Sun, and M. A. Saunders. Proximal Newton–Type Methods for Minimizing Composite Functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.
- [11] D.-H. Li, M. Fukushima, L. Qi, and N. Yamashita. Regularized newton methods for convex minimization problems with singular solutions. *Computational Optimization and Applications*, 28(2):131–147, 2004.

- [12] Z.-Q. Luo and P. Tseng. Error Bound and Convergence Analysis of Matrix Splitting Algorithms for the Affine Variational Inequality Problem. *SIAM Journal on Optimization*, 2(1):43–54, 1992.
- [13] Z.-Q. Luo and P. Tseng. On the Linear Convergence of Descent Methods for Convex Essentially Smooth Minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425, 1992.
- [14] Z.-Q. Luo and P. Tseng. Error Bounds and Convergence Analysis of Feasible Descent Methods: A General Approach. *Annals of Operations Research*, 46(1):157–178, 1993.
- [15] J. J. Moré. The Levenberg-Marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.
- [16] B. O’Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.
- [17] F. Oztoprak, J. Nocedal, S. Rennie, and P. A. Olsen. Newton-Like Methods for Sparse Inverse Covariance Estimation. In *Advances in Neural Information Processing Systems*, pages 755–763, 2012.
- [18] J.-S. Pang. A Posteriori Error Bounds for the Linearly-Constrained Variational Inequality Problem. *Mathematics of Operations Research*, 12(3):474–484, 1987.
- [19] J.-S. Pang. Error Bounds in Mathematical Programming. *Mathematical Programming*, 79(1-3):299–332, 1997.
- [20] H. Qi and D. Sun. A quadratically convergent newton method for computing the nearest correlation matrix. *SIAM journal on matrix analysis and applications*, 28(2):360–385, 2006.
- [21] S. Sardy, A. Antoniadis, and P. Tseng. Automatic Smoothing with Wavelets for a Wide Class of Distributions. *Journal of Computational and Graphical Statistics*, 13(2):399–421, 2004.
- [22] M. W. Schmidt, E. Berg, M. P. Friedlander, and K. P. Murphy. Optimizing Costly Functions with Simple Constraints: A Limited-Memory Projected Quasi-Newton Algorithm. In *International Conference on Artificial Intelligence and Statistics*, page None, 2009.
- [23] P. Tseng. Error Bounds and Superlinear Convergence Analysis of some Newton-Type Methods in Optimization. In *Nonlinear Optimization and Related Topics*, pages 445–462. Springer, 2000.
- [24] P. Tseng. Approximation Accuracy, Gradient Methods, and Error Bound for Structured Convex Optimization. *Mathematical Programming, Series B*, 125(2):263–295, 2010.
- [25] P. Tseng and S. Yun. A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization. *Mathematical Programming, Series B*, 117(1-2):387–423, 2009.
- [26] N. Yamashita and M. Fukushima. On the rate of convergence of the Levenberg-Marquardt method. In *Topics in numerical analysis*, pages 239–249. Springer, 2001.



- [27] I. E.-H. Yen, C.-J. Hsieh, P. K. Ravikumar, and I. S. Dhillon. Constant nullspace strong convexity and fast convergence of proximal methods under high-dimensional settings. In *Advances in Neural Information Processing Systems*, pages 1008–1016, 2014.
- [28] G.-X. Yuan, C.-H. Ho, and C.-J. Lin. An Improved GLMNET for L1-Regularized Logistic Regression. *The Journal of Machine Learning Research*, 13(1):1999–2030, 2012.
- [29] K. Zhong, I. E.-H. Yen, I. S. Dhillon, and P. K. Ravikumar. Proximal Quasi-Newton for Computationally Intensive  $\ell_1$ -Regularized  $M$ -Estimators. In *Advances in Neural Information Processing Systems*, pages 2375–2383, 2014.
- [30] Z. Zhou and A. M.-C. So. A Unified Approach to Error Bounds for Structured Convex Optimization Problems. *arXiv preprint arXiv:1512.03518*, 2015.
- [31] Z. Zhou, Q. Zhang, and A. M.-C. So.  $\ell_{1,p}$ -Norm Regularization: Error Bounds and Convergence Rate Analysis of First-Order Methods. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 1501–1510, 2015.